Review of Business and Economics Studies
2019, Vol. 7, No. 3, 25-36

DOI: 10.26794/2308-944X-2019-7-3-25-36

High-Frequency Trading in the Modern Market Microstructure: Opportunities and Threats*

Mikhail Zharikov

Doctor of Economics
World Economy and World Finance Department, Professor;
Institute for World Economy and International Finance Studies, Senior Scientific Fellow;
Financial University, Moscow, Russia
michaelzharikoff@gmail.com
http://orcid.org/0000–0002–2162–5056

Abstract

The article covers some ideas about the research on high-frequency trading and financial market design. The topic is time-relevant because today there exists a need to convince traders that there is a simple structural floor in the way that the financial markets are designed. The article reveals the significance of trading on the floor that the foremost fundamental constraint is limited time. The author proves that time on the financial market feels, to some extent, infinite when someone counts it in millions of seconds, but time is nevertheless finite. The author then gets into the actual research on high-frequency trading in the financial market design. The motivation for this project is to analyse activity among high-frequency trading firms by which investments of substantial sums of money are understood as economically trivial speed improvements. The theoretical significance of the research's outcomes lies in outlaying the systemic approach to dealing with stochastic control problems in the context of financial engineering. The practical relevance of the paper lies in the mechanism that allows solving problems surrounding optimal trading, market microstructure, high-frequency trading, etc. The article concludes by talking about the issues in the modern electronic markets and by giving lessons to dealing with them in the long run.

Keywords: financial engineering, financial innovations, high-frequency trading, world financial market
JEL Classification: F37

Introduction

It is worth going through some of the main features of the US equity markets today because they are quite different than they were five or ten years ago (Arner & Taylor, 2009).

First, markets are predominantly electronic. The trading happens on computers. Electronic trading has equity that dominates as the other primary mechanism of the exchange.

Second, there is an idea to think of the exchange as a mechanism for centralising trade by bringing buyers and sellers together, so that there are not search frictions. What has happened in the US in the past five years is that the opposite turn has occurred in that trading has become decentralised or fragmented (Beder, 2009). In particular, for various reasons, there is no one primary exchange. It used to be that for any specific stock which is traded either on NASDAQ or NYSE, but now there is a handful of them, and they are all important in a sense that each of those exchanges accounts for at least 5 per cent of equity trading. So, there are many venues, and the trade is no longer centralised (Chorev & Babb, 2009).

* The article is as part of the fundamental research titled The Formation of a Regime Favouring the Development of Digital Technologies conducted at the budgetary expense out of funds donated to the Financial University.

Most of the venues are organised as exchanges. They account for about 70 per cent of trade, and these exchanges are operated typically as electronic limit order books in the sense of an open market. People can submit orders to buy and sell, and they attach prices, and when prices cross, there is trade (Elyanov, 2009).

It is opposed to the dealer market or a specialist market, which is the way historically the New York Stock Exchange was organised. About 30 per cent of trade occurs on alternative kinds of venues. There are things like electronic crossing networks (ECNs), dark pools, internalisation, OTC market makers, etc.

Finally, the most striking feature is that the participants are increasingly automated. It used to be that if there was a hedge fund and there was a portfolio manager, and he/she wanted to buy a million dollars' worth of Google, there was also some entity who is a trader, and he/she knows how this sort of these things works (Griesgraber, 2009). Now computers do that. On the by-side, there are investors under the rubric of algorithmic trading who either themselves or on an agency basis present themselves as service providers by brokers. They will take large parallel orders and slice them, dice them over time and across exchanges, and then trade them.

Earlier the traders who were providing liquidity, the market makers, used to be human traders. Now in most of these markets, they often go to the rubric of high-frequency trading. One dominant kind of frequency trading is essentially providing liquidity and providing market-making services. Overall, these are all quite recent trends.

The interactions between an algorithmic trader and a high-frequency trader are challenging to predict. There was the famous flash crash of May 2010. The US Securities and Exchange Commission (SEC) reported that what happened was that in about five minutes, the market fell 5 per cent based on no news or fundamental information whatever. Then in the next five minutes, it recovered. It is a blip that came about from some pathological interaction between an algorithmic trader and high-frequency traders (Reinhart & Rogoff, 2011).

It raises two classes of essential questions. One is from the perspective of the system, policymakers, regulators, etc. who deal with issues such as: Is there a need in dark pools? Is it reasonable to have so many exchanges? Issues of class two come at the level of individual participants where there is no possibility to solve these decision problems.

If someone is trying to buy some stock, he/she has to decide whether they are going to use a dark pool or whether they are going to use an exchange? How is it possible to accomplish this?

There are two specific problems related to high-frequency trading in market microstructure.

The first is understanding the importance of latency. The second is understanding the role of dark pools in markets.

Latency is the delay between making a trading decision and its implementation. If someone decides to buy a hundred shares of Google, and he/she transmits that order to NASDAQ, how long it takes before that quantity is taken from the order book? Similarly, there is an order outstanding, and someone wants to cancel it. How long does it take between when someone makes that decision, and when those long orders are pulled from the matching engine and are no longer eligible for execution? It used to be the domain of IT-people, but in the past few years, it centred the public discussion. That is the idea of collocation (Yefremenko, 2007).

High-frequency traders often confound other investors by issuing and cancelling simultaneously. Maybe it is good to be able to trade quickly, although bullying does not sound so good. It is the technological arms race that separates winners and losers and how fast they can move.

Why is latency important? It is crucial to value the importance of low latency, and phrase it a different way: what is the cost associated with having latency?

Before 1980, it used to be that if someone put an order to buy a stock, it would take two minutes for that to occur. Later it came down to about 20 seconds. As of 2007 latency numbers came in hundreds of milliseconds. If someone is making trading decisions on that time scale, that is not humans, that are the computers that are trading with each other (Pisani-Ferry & Sapir, 2010).

In another couple of years, there will be the single-digit millisecond. So, if someone wants to send news from Chicago to New York, the speed of light limits them. That cannot happen faster than five milliseconds. If someone wants to be trading in less than a millisecond, that means there is a need for physical proximity.

The trend is to go even below that — a technology has driven it. Different exchanges emerged that offer technology benefits relative to income benefits. It is best to make personal decisions with the latest information possible. For example, if the traders are looking to sell 100 shares of Apple stock, the price that they are willing to sell at depends on the price other people are willing to sell or buy it. It depends on the price at different exchanges.

So, as traders digest more recent information, that will alter the price. There will be some advantage to having low latency. If there are two traders, and they are doing very similar trading strategies, typically the winner takes all. The fastest will get all the profits, and the other will get knocked out. That offers certain advantages. Depending on who they are, these different effects might kick in. How does investor benefit from having access to the latest information in terms of lowering costs? Here is a model. It can be called a stylised execution problem.

There is a trader who wishes to sell 100 shares over a very short time horizon, for example, 10 seconds. It is a problem that every trader faces at one level or another, and as was described earlier, the value that the trader perceives evolves, and the price at which the trader wants to get the share depends on this value. To figure how best to sell these 100 shares, the trader has to observe this valuation process. And if someone adds latency, that introduces a tracking error.

The trader does not precisely know what the value is. He/she only knew the value a millisecond ago. Because of that, he/she has to alter his/her actions, and that creates a cost, so latency becomes friction. What they do is they quantify the cost associated with latency. If they look at this execution problem and look at the transaction costs in the presence of latency, they will see how much worse that is. If they had known their latency in advance, they would normalise.

It depends on the volatility of the stock. The more volatile the stock is, the more critical the latency is. The more liquid the stock is, the more significant the latency is. That latency goes from being about 20 per cent of transaction costs to being 1 or 2 per cent. Does this make sense? Is this significant? It needs interpretation. For example, there is the stock, and the situation is normalised so that the bid offer was a penny. Most stocks in the US have a bid offer of a penny. What that thing is suggesting is that the value of decreasing latency from the human timescale to the machine timescale is about 20% of a penny or 20 mills. That seems a tiny number. But it is important to guess how much high-frequency traders make, people who have invested in being able to trade on this kind of timescale. Nobody really knows that, but self-reported numbers are of the same mode of magnitude (Mel'yantsev, 2015).

Earlier, where there was no ability to trade electronically, and the traders wanted to form it out, they wanted to pay an investment bank to trade for them and presumably they would have made that investment.

Latency is potentially crucial to all investors. It is a fundamental problem, the problem of selling 100 shares in 10 seconds. But how important it is, it depends on what the rest of the costs are. If they are at the most efficient cost level in terms of the commissions they have negotiated, latency is worth about as much. On the other hand, if they are retail investors, they are not paying five mills per share traded, they are paying 10 dollars to each trader. Those orders in magnitude are more than any of these. So, from their perspective, for a retail investor, this does not matter. The commissions and other things they are paying dominate the value of latency (Lane, Milesi-Ferretti, 2011).

Dark Pools

Dark pools are an alternative trade mechanism. If one thinks about a limit order book they want to buy, typically there is an offer price which is higher than the price at which they could sell, which is a bid price. There is a bid offer spread, a limit order book or an exchange. Someone is providing liquidity who may be a high-frequency trader, and they are going to charge for that, and this bid offer spread is what they charge.

What is an alternative mechanism? The idea of a dark pool is instead of having these intermediaries posting orders, i.e., people who directly trade with each other, there is just an anonymous pool where some people can declare they want to buy, some people can declare they wish to sell, and if there is a match, they will be matched with each other, and it will occur at the mid-market. It means no transaction costs. If the traders are trying to buy on an exchange with a market order, they are going to execute for sure. If they put an order into a dark pool, they will get it at a better price. It is a trade-off that occurs in many markets. It is a trade-off between uncertain trader at a better price, i.e. the dark pool, or guaranteed trade at a worse price.

For example, in an eBay auction, the people typically can pay a price premium, get the item they want with certainty, or they can participate in the auction, get it cheaper, or they will not get it (Kose, Prassad, Rogoff, & Wie, 2009).

What is very specific about it here is a simple stylised model in a financial context where investors have two options. One is a guaranteed market where someone can trade with certainty, but they pay a transaction cost, they pay the bid-offer spread. It is a dealer market or electronic order book. The second option is a dark pool. Here someone puts the order into one of these electronic crossing networks. If a trade occurs, it occurs at mid-market. It has zero transaction cost. But they are not sure it is going to happen. So, there is a need to evaluate these two alternatives. The key ingredient in this model is information ladders. It might be important for a couple of reasons.

First of all, if the stock is going up, and someone is pretty sure that the stock is going up, that is going to affect whether someone wants to trade with certainty, or they will be uncertain. For example, if they are confident, they are willing to pay the transaction costs, and they want to trade with certainty. The information matters here. However, other people's information matters also, because when they trade, they are trading with others in the case of a dark pool. If they are systematically trading with people who have more information than the other traders, maybe that is not going to work out for them so well in the end. One critical thing is going to be modelling information. There is a need to have a model where there are three kinds of traders or speculators. Everybody observes a signal about what is going to happen with the price. The speculators are just trying to make money of the price swings. On the other hand, they have intrinsic buyers and sellers. They also would like to buy low and sell high. However, they have their reasons to trade. They have idiosyncratic desire to trade (Lebedeva, 2013).

There is an equilibrium in this model, and there are some predictions. For example, the more information traders have, the more they are willing to go to the guaranteed marketplace. If they are very well informed, and they know the price is going to go up, they want to buy with certainty. On the other hand, if they are less informed if they have little or no information, they are trading only on idiosyncratic reasons.

The traders are willing to trade in a dark pool; they are eager to take that risk as it does not make sense for them to pay transaction costs. If someone imagines a world with a dark pool and a world without the dark pool, transaction costs will be higher in the presence of a dark pool. Where are these transaction costs coming from? It is a bid offer spread which is going to be set by market-makers, who are trying to make money. If they have a dark pool present, those market-makers are going to end up systematically losing more money. They are going to widen their spreads to compensate for that.

The presence of a dark pool is going to deteriorate the quality of the guaranteed market. Investors in the dark pool are going to experience adverse selection. If someone trades in the guaranteed market, no matter whether the price is going up or down, they are going to get that share. If they trade in the dark pool, they are not sure, if they will or not, but if they are trying to buy, what will happen is typically when the market is going down, their order will get filled, and when the market is going up, it will not (Khmelevskaya, 2015).

So, precisely in the circumstances where the people do not want to trade because they could have bought it cheaper later, they will trade and otherwise, they will not. A naïve person will look at dark pools and say there are no transaction fees; they are trading on the mid-market. Because they are not trading for sure, and because their trades are going to be correlated with what happens to price afterwards, they are paying an adverse selection fee. It is implicit. It is not explicit like the bid-offer spread. Statistically, they are paying this fee. It can be of the same order of magnitude as the bid-offer spread. So, the dark pool is not as good as it looks, because it decreases welfare (Jorda, Schularick & Taylor, 2009).

The Technology of High-Frequency Trading (HFT)

In 2010, Spread Networks Co. invested 3 million dollars in digging a high-speed fibre optic cable connecting financial markets in New York City to financial markets in Chicago. The failing feature of this cable is that it was dug in a relatively straight line. The straightness of this line shaved round ship data transmission time by three milliseconds, by 3000th of a second. To put that into context, blinking an eye takes several hundred milliseconds (Kemenyuk, 2009).

Economists have been working on this project for over three and a half years, which is more than a hundred billion milliseconds. Three milliseconds do not sound much. Industry observers described it as an eternity. The joke at the time was that the next innovation would be to dig a tunnel, go through the earth, right around the earth because that will further shave data transmission time. This joke materialised. The spread cable is already obsolete. It is not a tunnel through the earth, but because light travels faster through the air than through fibre optic cable, there is special relativity that one should be talking about.

Micro-waves have further shaved data transmission time. The time is now down to eight-and-a-half milliseconds. The Einstein bound is eight milliseconds round chip between these two markets. Other races are occurring throughout the financial system, sometimes measured as finally as millions or even billions of seconds (Helleiner, 2009). In the last few months alone, the most recent innovation has been laser beams which have the speed properties of microwaves, but more reliable in bad weather. There have been announcements to do with the release of public information early by firms like BusinessWire.

In this project, the arms race is looked at from the perspective of market design. It is an academic approach which takes for granted that participants in the market act rationally and in their self-interest concerning market rules. They take seriously the possibility that the rules themselves are flooded.

HFT is a rational optimising concerning market rules. Are the market rules themselves optimal? At a deeper level, the question is: what is it about a market design that induces the arms-race-like behaviour in this design system? The central point is going to be that the arms race among high-frequency trading firms is a symptom of a simple structural floor in market design. This floor is continuous-time trading.

Continuous-Time trading means that people can trade, buy themselves stock or buy themselves futures contracts or buy themselves anything else at literally any instant during the trading day or instant measured as finally as computers allow. What the industry is going to propose as an alternative is to make time discrete. More specifically, the industry is going to suggest replacing the continuous-time limit order book market (Dorrucci & McKay, 2011).

The order books are the predominant market design used by financial exchanges today with discrete-time, which can be called frequent badge auctions. These are the uniform price double auctions conducted very frequently at a discrete-time throughout the day. It is a massive document, a massive argument with four parts. The first thing worth showing is some empirical evidence that continuous markets do not work as they are expected to work at very high-frequency time scales.

There are two ways to quantify this phenomenon. One is to ask what the correlation exists between these assets at different time horizons? At horizons of a few seconds or more, the correlation is essential. They should be perfectly correlated. They track the same index, but at a millisecond, the correlation is less than 0.01 (Chen, Milesi-Ferretti, & Tressel, 2012).

If someone takes the perspective of a Chicago observer treating New York as the recent past or a New York observer treating Chicago as the recent past, or if special relativity is just ignored altogether, the correlation completely breaks down at high enough frequency as one. Then the second way to quantify this phenomenon is to ask how often this breakdown in pricing relationships creates free money. Does it create technical arbitrage opportunities? What can be seen here at this moment is buying cheap in New York and selling expensive in Chicago, which is an essentially riskless profit opportunity. There are in order of 800 such opportunities per day.

The typical intuition about arbitrage, especially about obvious arbitrage opportunities is that almost the same security is trading in two different markets at two different prices. The intuition about obvious arbitrage opportunities is that they get competed away. The first thing is to look at the duration of these technical arbitrage opportunities throughout the data.

Each day has 23.4 million milliseconds. The prices should move together, but the prices get out of track as well, and money can be made buying the cheap one, selling the expensive one. The duration of these arbitrage opportunities has come down over time. In 2005 they lasted on the order of 100 milliseconds on average. By 2011, it was sub-10 milliseconds, i.e. less than one-one-hundredth of a second. If there is a fifty-millisecond arbitrageur in 2005, that is whenever prices were dislocated for at least 50 milliseconds; they could get that arbitrage. In 2005 they were state-of-the-art and got almost everything. By 2011, they are entirely obsolete, and traders get virtually nothing.

Durations have come down overtime. But profitability per arbitrage opportunity has remained relatively constant. There is nothing in the market design that allows prices to move at the same time. The people still make just as much money per arbitrage opportunity at the end of the data as at the beginning of the data. The exceptions include a blip-up during the financial crisis in 2008 (Afontsev, 2014). And the bigger the price movements, the more common are significant price dislocations.

2008 was the best year in history to be a high-frequency trader. Profits were lower pre-2008 and also post-2008.

Frequency does change overtime very substantially, but it is explained almost perfectly by just asking how volatile the market was on a particular day. It is a complementary way of looking at the same phenomenon as this.

If one looks though, at one millisecond, one sees that in all years at high enough frequency correlations entirely fall apart, and again there is nothing in the market design that allows security prices to move at the same time. These results suggest that the arms race should be viewed as something of a constant of the market design rather than a profit opportunity that gets competed away overtime. These correlation curves show that competition overtime, i.e. the high-frequency speed race. It increases the speed which information makes it from one security's price into another security's price but does not eliminate the underlying phenomenon that prices cannot move at the same time.

Correlations break down at high enough frequency. In the technical arbitrage, a way to interpret that is that competition does increase the speed requirements for capturing arbitrage opportunities. It raises the bar, but competition cannot eliminate the arbitrage opportunity, it does not reduce their size, and it does not affect their frequency. Frequency is affected by volatility, but not by speed competition *per se*. These facts are going to be taken to inform and then also explain the theoretical model (Andronova, 2012).

The thing one really would not like to emphasise is that it suggests the tip of the iceberg in the race for speed. There are hundreds of pairs of securities that are very similar to the S&P 500 pairs or highly correlated. In fragment to equity markets, there are even simpler trading opportunities. The same stock can take as an example that trades on thirteen different exchanges and fifty different dark pools. All of those prices should move exactly together, but nothing in market structure allows them to run at the same time, and one can make money buying the cheap and selling the expensive one (Titarenko & Petrovskiy, 2015).

Correlations that are high but far from 1 can be exploited in a statistical sense. There is a race to get to the top of the order book, which is an artefact of fat tick sizes. It shows up in a lot of contracts that trade on the Chicago Mercantile Exchange, which has fat ticks. Interestingly, there is a discussion in Washington about mandating a larger tick size in US equities with a theory that will invigorate the market for small cap stocks and equity research. From the frequency perspective widening the tick's scope exacerbates the race to get to the top of the book. There is a race to respond to public news like a Fed announcement at 2 p.m. When the Michigan Consumer Confidence number comes at 10 a.m., there is also a race to react. There is no need to try to put a precise estimate of the total prize in the race, but common sense suggests a lot of money is on the line (Semedov, 2015).

The Model of High-Frequency Trading

The theory models are going to do two things: one is going to be a critique of continuous trading. And the second is going to help articulate what exactly is wrong with prepetition based on the speed. What are precisely the economic consequences of this race for speed?

It is a very simplified model which tries to help explain the facts and then serve these two related purposes. At the core, it is a straightforward model. It makes some crucial points. There is a security X that trades on a continuous order book market. There is a publicly observable signal Y of the value of security X. There is a need to make a purposely very strong assumption that security X is perfectly correlated to public signal Y. Moreover, at any moment in time, one can causelessly liquidate X and get Y.

There is a best-case scenario for price discovery and liquidity provision in a continuous-time market. There is a model in which it should be economically trivial to provide liquidity in the market for X. This model does not have asymmetrical information. It does not have inventory costs. It does not have the usual sources of costly liquidity provision. X and Y are understood as a matter for four pairs of securities that are highly correlated.

Signal Y involves the compound Poisson jump process. There are two types of participants in the model: investors and trading firms. Investors represent N users of financial markets, i.e. mutual funds, pension funds, hedge funds, etc. They arrive randomly to market and, needing to buy one unit of this security X; they have a random arrival rate lambda invest. It is equally likely that they need to buy a unit versus selling a unit. They are very mechanical. They trade at market immediately upon arrival.

The other participants in the model are trading firms or equivalently high-frequency traders, algorithmic traders, etc. They do not have an intrinsic demand to buy or sell X. They are traders that want to buy low and sell high. Their goal is to maximise profits per unit time. The number of trading firms is exogenous. There is capital coming in, trading firms present in the market, and then later the entries will be endogenised or allowed for costly entry by investing in a speed technology.

Latency should be taken into account at first in a straightforward and stylised way, again towards building up the best case for the performance of a continuous market. Initially, there is no latency in observing why. So, whenever this security, that whenever this single Y jumps around, everybody in the game sees it immediately for free.

Moreover, there is no latency in submitting orders to the exchange. If someone decides at some time that they want to send an order to the exchange to buy, their order reaches the exchange at precisely that time. If the order to buy and the order to sell reaches the exchange at the same time, the exchange processes these two requests one at a time. It is called serial processing.

Part of the best-case scenario is assuming away latency. Given this model set-up, there is no asymmetric information. There are no inventory costs. Everybody's risk is neutral. That complication is going to lead to a healthy outcome which should be economically trivial to provide liquidity in the market for X. But that is not what happens in a continuous limit order book, due to a phenomenon which can be called sniping.

Suppose single Y jumps from Y lower bar to Y upper bar. The price in Chicago of the e-money futures goes up two ticks. It is the moment in which the correlation between Y and X temporarily breaks down. The world changed. The market jumped a couple of ticks. The quotes are now incorrect. The traders send a message to cancel old quotes and replace them with new quotes based on the new public information.

But at the same time, other trading firms try to snipe the still quotes. Again, the world is changing. Someone sends a message to cancel these quotes. At the precise same time, all of the traders and the other trading firms send a message to buy at the ask price. The ask is too low relative to the information. Since continuous markets process these requests in serials, one at a time, in order of arrival, it is possible that one of the requests to trade at this old price is going to reach the exchange before the request to cancel these quotes and replace them with new quotes based on the latest information. It is not only possible but probable because one of the traders tries to cancel and everybody else attempts to exploit the still quotes. It is an asymmetry.

When there is a big jump, liquidity providers get sniped with high probability N minus one over N. If there are 100 trading firms, it is 99 out of 100 chances that they get sniped when there is a big jump. In a continuous market, there is symmetrically observed public information, and these jumps in Y create technical arbitrage opportunities.

Everybody understands equally well that X is worthwhile. Y jumps at the same time, and yet somebody is going to make money from the first message process to trade at the old price. It is not supposed to exist in an efficient market. There should not be such simple arbitrage opportunities in an adequately designed market, and it is closely associated with the correlation breakdown phenomenon. The reason for this is that equilibrium, the cost of getting picked off by all of the traders are being passed on to investors. It is a cost of doing business. It is a cost of liquidity provision. The traders get an equilibrium in which N trading firms provide liquidity to real investors. It provides bids and asks.

Trading firms are going to be indifferent between these two roles and equilibrium, and in practice, most of the high-frequency trading firms perform both roles, mutually throughout the day. There are some exceptions to that. The difference between the price at which one is going to buy and the price at which one is going to sell is going to have to compensate for the risk of getting picked off by all of the traders. If someone works through math, they get an equation that describes the bid-ask spread, which creates revenue for a trader from investors. Investors come along and pay the bid-ask spread. That is good news for the trader as a liquidity provider. That compensates for the risk of getting sniped by all of the traders.

As a subtle economic interpretation on the left-hand side of this equation is the revenue from investors due to a non-zero bid-ask spread. On the right-hand side of this equation, there are the rents to trading firms from this technical arbitrage that are caused by the market design. What happens if so far investors show up wanting to buy themselves one unit? Or what happens if investors show up wanting to buy themselves two units? Or want to buy or sell a million shares? If someone is a liquidity provider, and they provide a profound order book, they provide a million shares at the bid and a million shares at the ask.

There is a jump where the traders are all going to try to pick one off for all million shares because it is free money for all of them times a million. The costs of providing a deep book scale correlate linearly with how deep of a book someone provides. But not all investors want to buy themselves a million shares. Many want to buy themselves just a hundred shares. The benefits of providing a deep book do not scale. This sniping cost causes not only a non-zero bid-ask spread; it also causes markets to be unnecessarily thin. One is not going to be able to provide a million-share book; rather, they are going to charge a considerable price for quoting that much depth because they are worried about getting picked off.

The next thing they do in the model is endogenising entry. So far, it is free to observe innovations in Y, and there is just some exogenous number of trading firms. There are a hundred trading firms in the market. Now everybody can observe innovations in Y at slow latency for free, at the latency of delta slow. They can pay a cost, a speed cost to observe innovations in Y faster. It is going from the slow cable to spread networks cable. They are going from the spread networks cable to microwaves or from 2012 microwaves to 2014 microwaves.

In equilibrium, the traders get a very similar structure — everybody snipes. Fast-raters are indifferent between the two roles. When one works through the math, they are going to skip these equations. They get a subtle characterisation of equilibrium where the total revenue from investors that the liquidity provider earns equals the total expenditure on speed by high-frequency trading firms.

Continuous trading creates these technical arbitrage opportunities, say, 20 billion dollars a year. And then high-frequency trading firms invest real resources of three million dollars cables.

In the equilibrium of the model, all the rents from these technical arbitrage opportunities get soaked up and get dissipated in competition to realise arbitrage.

There are equivalents in the arbitrages between prize and the speed race, the amount expanded on this speed race and the end cost to real investors. One should always keep in mind that profits in financial markets have to come from somewhere. In the model, they are coming from end investors.

The model points to two market failures. One is the phenomenon called sniping. Technical arbitrage opportunities are simply embedded in the design of continuous limit order book market. These are opportunities that should not exist in the efficient market and allow earning rents from symmetrically observed public information.

Everyone has intuition. If someone is a hedge-fund analyst, and he/she figures out something about a company that nobody else in the market knows, they can make money from that. But in this model, one can figure out something that the rest of the market knows. The traders can make money from that. A second market failure is that this free money creates a speed race. Mathematically for those traders, this is a prisoners' dilemma.

The arms race in the model is constant. Nothing in the analysis depends on whether the difference between fast-raters and slow-traders is seconds or milliseconds or microseconds and nanoseconds. Instead, this sniping phenomenon is an equilibrium feature of continuous trading. It does not get competed away.

The model encourages a constructive way of thinking about high-frequency trading firms. In the model HFTs endogenously decide to perform two roles: a useful role and a negative role. The useful role is in providing liquidity and enhancing price discovery for real investors. It is providing liquidity in the market for X.

The negative role is sniping still-quotes. When the market changes picking off old prices, these sniping still-quotes look like zero-sum HFT. Frequent batch auctions preserve the useful function, the price discovery and liquidity provision but eliminate the rent-seeking function.

Markets today are more liquid than they were in the pre-HFT era. It is a lot cheaper to trade in 2014 than it was in the 1990s. But there is a vast information technology revolution as financial markets switched from humans toward electronic bases. All gains were realised in the relatively early stages of the IT-revolution.

Conclusion

The take-away from the empirical record is that information technology has been unambiguously good for markets, but there is no evidence that the speed race has been good for markets. The research suggests that the speed race has been negative, at least in recent years, at a millisecond or microsecond level.

There is an alternative to continuous trading which can be called frequent batch auctions. At a high level, frequent batch auctions are very analogous to current practice to continuous limit order book trading with the vital exception that time is discrete. Stocks trade in a penny, meaning that it is a discrete price increment. The trader is not allowed to bid a millionth of a penny more than the other one is to jump ahead in the queue. There is a discrete price increment.

Discrete time necessitates batch processing.

The proposal is to divide the day into equal intervals, say, a hundred milliseconds. During this interval, traders submit bids and asks. The same language is currently of a price quantity in a direction. Orders can be freely cancelled, withdrawn and modified any moment in time. At the end of each interval, the exchange batches together.

Supply and demand either do not cross or they do — if they do not cross, then there is no trade. All orders remain outstanding for the next batch interval. Most stocks have no trading activity. Instead, what one can see as a market participant are a supply and demand.

The other case is that supply and demand do cross at some price, say p*, in which case the logic is one of a uniformed price auction. In the 1960s, it was initially proposed by Milton Friedman and adopted in the 1990s by the US Treasury for the US Treasury market auctions.

If a trader bids more than p* or higher or someone else asks lower, one transacts the full quantity at p* at, say, 10 dollars, at the uniformed price of the auction. If one bids precisely 10 dollars, one might get rationed.

The suggested rationing rule is to respect time priority. If the order has been sitting in a book for ten seconds and the other order is new to the book in this trading interval, the previous order has precedents of the latter. But if the first order and the second order reach the exchange at the same interval, they can be treated equally.

The market clears at price p*. The auction is very similar to continuous limit order book trading. After the auction is computed, the price is announced, the quantity of the supply and demand curves are announced, and there are some more details about the information policy. Why is a frequent batch auction an attractive alternative to continuous trading? There are two reasons for that. The first, the apparent reason is that frequent batching reduces the value of a timing speed advantage. In the discrete time, the market is trading one per second, and one trader is a hundred millionth of a second faster than the other one.

The second, more obscure reason why frequent batching is attractive is that it transforms competition on speed into competition on price and eliminates the sniping phenomenon. Suppose someone is trying to provide liquidity. There is a jump in the public signal Y, so there are either many jumps or the Fed makes an announcement at 2 p.m. What is evident to all participants is that the market is going to tick up several points. In the continuous market, there is a race to pick up still-quotes at these incorrect prices. In the batch market, there is an auction.

The trader gets the buy, the other one pays, and the former will be able to make the rent. Pure speed competition is designed away. Competition is based on price. In equilibrium, the benefits of frequent batching relative to continuous trading eliminates sniping, which enhances liquidity. Narrower bid/ask spreads go in greater depth. It stops a socially wasteful arms race.

The cost is that investors have to wait to transact.

The equilibrium analysis focuses on liquidity and a socially wasteful arms race. Another case for discrete time trading is based on computational advantages. Continuous time trading implicitly assumes that computers and communications technology are infinitely fast. If an event happens on the NYSE, the NASDAQ knows about that immediately.

All of this is a manifestation of the fact that the information does not travel instantaneously fast. It takes a few hundred microseconds or a couple of milliseconds to get from one place to the other. Discrete time respects these computational and communications limits. For an algorithmic trader, discrete time means one sees what happens on the market at time t. One has a block of time to think about it, or for the algorithm to think about it and make decisions at t+1, and then one sees what happened at t+1 or t+2.

Programming is a clean environment, whereas, in the continuous market, one does not know what they are going to learn. One also does not know what information others in the market have.

For exchanges, continuous trading creates a computationally impossible task. Exchanges invariably get back-logged. If there is a lot of activity exchanges, it takes time to process that activity. In discrete time the computation is trivial.

For a regulator in a continuous market, it is difficult to parse the audit trail. It takes months for regulators to figure out what triggered the flash crash, and even today, the understanding of that day's events is far from complete. There is a discrete time of the simple audit trail. It happens at t or t+1. Once per hundred milliseconds is very different in terms of the audit trail which yields from once per nanosecond. A nanosecond is too small relative to noise in communications and computing time.

There have been multiple other policy responses to the high-frequency arms race, for example, the Tobin tax.

To summarise, one looks at the arms race in the perspective of market design. The traders do not think the root problem is evil high-frequency trading firms. First of all, continuous time markets are in fiction. Correlations break down. There are frequent technical arbitrage opportunities. Second, these technical arbitrage opportunities induce a never-ending speed race, which looks like a constant of the design. The bar gets higher each year, but it does not compete away.

The theoretical model shows that the root causes market design. There is continuous limit order book trading. The arms race is an equilibrium feature of the design. It harms liquidity and is socially wasteful.

The research shows that frequent batch auctions are an attractive market design response and an equilibrium that eliminates sniping. It stops the arms race and enhances liquidity and has computational advantages. The costs of that are that investors have to wait to trade.

There are two essential parameters in the model: tick size in time and tick size in price. The question is how a regulator or exchange owner would optimally choose those parameters. The model assumes away that tick size in price. It treats the tick size in prices being arbitrarily fine. And this is indiscrete time auction.

There is a clear need to have a discrete tick size in continuous markets. There is an economic model of what the optimal tick size is. A finer tick size should be chosen, which means more accurate price discovery, although one does not have a concrete way of thinking about it.

In the end, Milton Friedman once said, *"…it is really hard to change policy, there is enormous inertia in the private sector and especially regulatory arrangements as an inertia of the status quo. And it is only upon a crisis that there is real change."* And he also said, *"…our job as economists is to develop good ideas and then to keep them active, keep them part of public discussions. So, that when a crisis hits, people reach for a good idea rather than reaching for a lousy idea."*

References

Afontsev, S. A. (2014). Dominirovaniye dollara: est' li al'ternativy [Dollar's dominance: Are there any alternatives?]. *Rossiya v global'noi politike*, *4(12)*, 120–129.

Anrdonova, N. E. (2012). Osnovniye predposilki i perspektivy konvertiruyemosti rublya i formirovaniye mezhdunarodnogo finansovogo tsentra v Rossiyskoi Federatsii [Basic causes and prospects of the rouble's convertibility and the formation of the international financial centre in the Russian Federation]. *An Economic Review of the Republic of Tatarstan*, *4*, 5–10.

Arner, D. W., & Taylor, M. W. (2009). The Global Financial Crisis and the Financial Stability Board: Hardening the Soft Law of International Financial Regulation. *University of New South Wales Law Journal, 32(2)*, p. 489.

Beder, S. (2009). Neoliberalism and the Global Financial Crisis. *Social Alternatives*, *28(1)*, p. 18.

Chen, R., Milesi-Ferretti, G. M., Tressel, T. (2012). Euro Area Debtor Countries: External Imbalances in the Euro Area. *IMF Working Paper*, 2012, *12(236)*, 1–22.

Chorev, N., & Babb, S. (2009). The Crisis of Neoliberalism and the Future of International Institutions: A Comparison of the IMF and the WTO. *Theory and Society, 38(5)*, 459–484.

Dorrucci, E., & McKay, J. (2011). The international monetary system after the financial crisis. *European Central Bank Occasional Paper Series*, *123*, p. 10.

Elyanov, A. Ya. (2009). Mirovoi ekonomicheskiy krizis i razvivayushchiyesya strany [World economic crisis and emerging economies]. *Mirovaya ekonomika i mezhdunarodniye otnosheniya*, *10*, 24–32.

Griesgraber, J. M. (2009). Reforms for Major New Roles of the International Monetary Fund? The IMF Post-G-20 Summit. *Global Governance*, *15(2)*, p. 179.

Helleiner, E. (2009). Special Forum: Crisis and the Future of Global Financial Governance. *Global Governance*, *15(1)*, p. 1.

Jordà, Ò., Schularick, M., & Taylor, A.M. (2011). Financial crises, credit booms, and external imbalances: 140 years of lessons. *IMF Economic Review*, *59(2)*, 340–378.

Kemenyuk, V.A. (2009). Poryadok posle krizisa: kakim yemu byt'? [An order after the crisis: how should it look like?]. *Mezhdunarodniye protsessi*, 3(7), 1.

Khmelevskaya N. G. (2015). Prioritety vneshnetorgovoi politiki Rossii v orbite ekonomicheskogo sotrudnichestva BRIKS [The priorities of foreign trade policy of Russia in the orbit of economic cooperation of the BRICS]. *Economic policy*, 2(10), 93–109.

Kose, M. A., Prasad, E., Rogoff, K., Wie, S.-J. (2009). Financial globalization: A reappraisal. *IMF Staff Papers*, 1(56), 8–62.

Lane, P. R., Milesi-Ferretti, G. M. (2011). External Adjustment and the Global Crisis. *IMF Working Paper*, 11(197), 3–18.

Lebedeva, M., M. (2013). Aktory sovremennoi mirovoi politiki: trendy razvitiya [The actors of the modern world politics: development trends]. *Review of MGIMO University,* 28(1), 38–42.

Mel'yantsev, V., A. (2015). Sravnitel'niy analiz modeley razvitiya Kitaya, Rossii i Zapada [The comparative analysis of the development models in China, Russia and the West]. *Mir peremen*, 3, 177–180.

Pisani-Ferry, J., Sapir, A. (2010). Banking Crisis Management in the EU: An Early Assessment. *Economic Policy*, 25, 341–373.

Reinhart, C. M., Rogoff, K. S. (2011). From Financial Crash to Debt Crisis. *American Economic Review*, 101(5), 1676–1706.

Semedov, S., A. (2015). Mesto Rossii v mirovoi politike i ekonomike [The place of Russia in world politics and economy]. *Aktual'niye voprosi innovatsionnoi ekonomiki,* 9, 9–10.

Titarenko M. L., Petrovskiy V. Ye. (2015). Rossiya, Kitai i noviy mirovoi poryadok [Russia, China and the new world order]. *Mezhdunarodnaya zhizn',* 3, 23–43

Yefremenko, I.N. (2007). Osnovniye napravleniya transformatsii mirovoi finansovoi arkhitekturi v usloviyakh finansovoi globalizatsii [Main directions of the world financial architecture's transformation in the conditions of financial globalization]. *Finasi i kredit*, *41(281)*, 43–51.

Высокочастотная торговля в современной микроструктуре финансового рынка: возможности и угрозы

Михаил Жариков

доктор экономических наук, доцент, профессор Департамента мировой экономики и мировых финансов, главный научный сотрудник Института мировой экономики и международных финансов, Финансовый университет, Москва, Россия
michaelzharikoff@gmail.com
http://orcid.org/0000–0002–2162–5056

Аннотация. В статье описаны основные положения исследования высокочастотной торговли и организации финансового рынка. Тема актуальна в связи с тем, что в настоящее время существует необходимость создания инструментов для участия в торговле упрощенными структурными продуктами, составляющими структуру рынка. В статье выявлена значимость биржевой торговли в ее зависимости от фундаментальной ограниченности времени на финансовом рынке. Автор доказывает, что время на финансовом рынке имеет иное измерение и по внешней видимости обладает бесконечностью, поскольку исчисляется в миллионах секунд, но при этом остается редким ресурсом. Впоследствии автор переходит к основному исследованию высокочастотной торговли в структуре финансового рынка. Цель статьи — проанализировать деятельность компаний, занятых в сфере высокочастотной торговли, которые осуществляют масштабные инвестиции в усовершенствование методов функционирования в жестких временных рамках. Теоретическая значимость результатов исследования заключается в изложении системного подхода к решению проблем стохастического характера в контексте финансового инжиниринга. Практическая значимость статьи состоит в разработке механизма, который позволяет решать проблемы, связанные с оптимальной торговлей, микроструктурой рынка, высокочастотной торговлей и др. В заключении автор систематизирует уроки из опыта современной электронной торговли на финансовом рынке и их решения в долгосрочном периоде.

Ключевые слова: финансовый инжиниринг; финансовые инновации; высокочастотная торговля; мировой финансовый рынок
JEL Classification: F37

Статья публикуется в рамках фундаментальной научно-исследовательской работы по теме «Формирование благоприятного режима развития цифровых технологий» по результатам исследований, выполненных за счет бюджетных средств по государственному заданию Финуниверситету.